READ
RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D6.4
Basic Layout Analysis

Markus Diem, Stefan Fiel, Florian Kleber
CVL

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	31.12.2016
Actual date of delivery	28.11.2016
Date of last update	21.12.2016
Deliverable number	D6.4
Deliverable title	Basic Layout Analysis
Type	report
Status & version	in progress
Contributing WP(s)	WP5
Responsible beneficiary	NCSR
Other contributors	CVL
Internal reviewers	URO,NCSR
Author(s)	Markus Diem, Stefan Fiel, Florian Kleber
EC project officer	Martin MAJEK
Keywords	Layout Analysis, Baseline Detection

Contents

1 Executive Summary

The basic layout analysis module extracts visual features from document images. These features include page segmentation (paragraphs), text-line segmentation, and the recognition of supplemental elements (e.g. images, ...). The current deliverable D6.4 is trained for images, handwritten/printed text and noise as supplemental elements. However, the implemented classification can be trained for any entity based on the requirements of e.g. certain collections (e.g. initials). A detailed evaluation of the basic layout analysis on competition datasets will be presented in D6.5. The module is part of the CVL READ framework. It is Open Source under LGPLv3 and available at github[1]. In addition to the command line testing routines, a plugin[2] for nomacs[3] is provided which allows for training and testing on either single images or a batch of images.

2 Super-Pixel

Document elements such as characters, words, or decorations need to be segmented in order to group and/or recognize them. Binarization algorithms such as Otsu or Su [1] are typically utilized for segmentation. This approach has a major drawback: one cannot segment black characters on white and white characters on black at the same time. This is why MSER [2] is used in order to extract connected components which we call *super-pixel*. Even though MSER overcomes the previously mentioned issue, it segments words rather than characters if cursive handwriting is present. That is why we build a scale-space with increasing circular erosions. Figure 1 shows the improvement of the proposed MSER extraction. For memory efficiency, each MSER region is approximated by an ellipse which is estimated by means of a PCA.

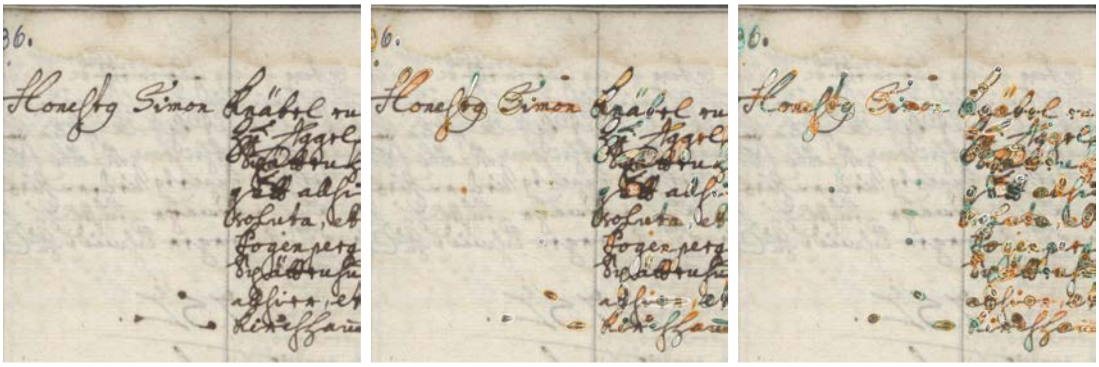


Figure 1: Detail of *_M_Aigen_am_Inn_007_0336_* (left), MSER output (middle), improved MSER output using erosions (right).

[1] https://github.com/TUWien/ReadFramework

[2] https://github.com/TUWien/ReadModules

[3] https://nomacs.org

3 Local Orientation

Text-lines may have changing local orientations because of perspective distortions, warped pages or simply because they are written at different angles. We use a local orientation estimation which is strongly related to that proposed by Il Koo [3]. A $N \times N$ neighborhood is extracted for each super-pixel. Then, projection profiles using the super-pixels' center-of-mass are extracted at different angles. The projection profiles are further transformed using the Discrete Fourier Transform (DFT). The DFT has large peaks if recurring frequencies are present. In general, text-lines have a (more or less) fixed line spacing which results in such a peak. Hence, the histogram with the largest peak indicates a super-pixels' local orientation. Since local orientations are similar with respect to the location, the orientations are smoothed using a multi-label graph-cut.

Figure 2 shows the local orientation estimation of a super-pixel. The orange histograms show the pixel's projection profiles sampled at eight different orientations. For illustration reasons only eight orientations are sampled, usually 32 projection profiles are created which results in an angular resolution of 5.6 °. The right image shows the DFT of the projection profiles. It can be seen, that the histogram at 90 ° has repeating peaks which result in a large peak after applying DFT. By this peak, we can determine the local orientation and the line spacing.

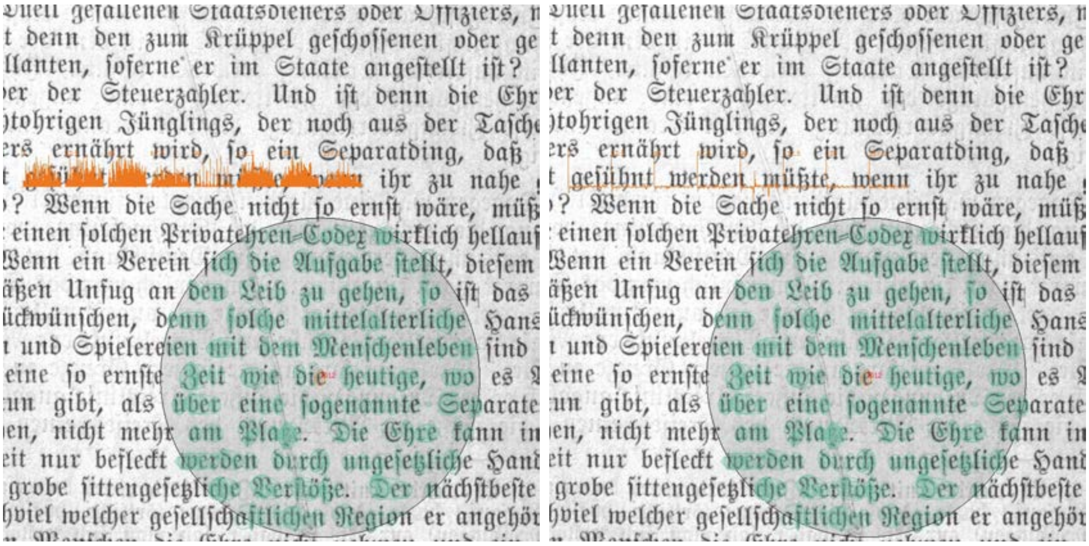


Figure 2: Projection profiles of a Superpixel (left), DFT (right).

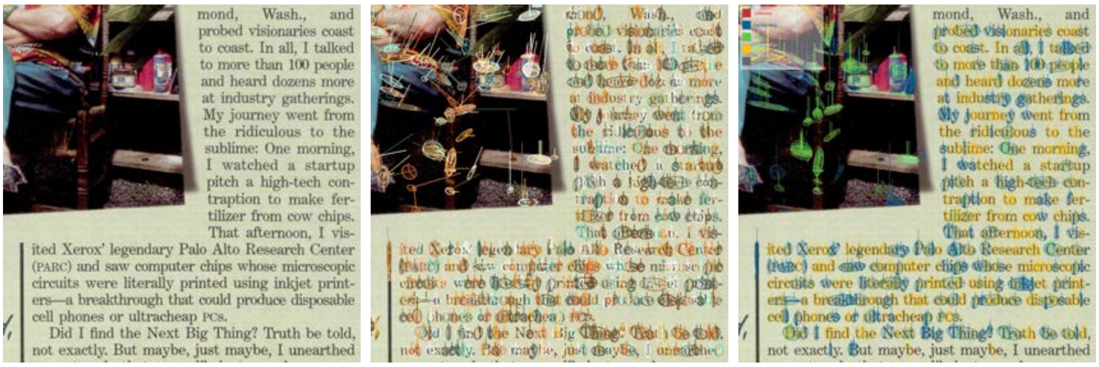
4 Component Labeling

Having extracted super-pixels with Maximally Stable Extremal Regions (MSER), the location and scale of Connected Components (CCs) are known. However, it is not known what elements are being found. Therefore, super-pixels are classified. We use local features - namely Oriented FAST and Rotated BRIEF (ORB) - in order to represent the local structure. ORB can be replaced by any other descriptor such as SIFT or DAISY.

ORB are chosen since they are computationally efficient and comparably compact (32 bytes). For classification, a Random Forest [4] is utilized. Again, any other machine learning (e.g. Support Vector Machines (SVM), Neural Nets) can be used. Random Forests are convenient for their fast training, multi-label classification, and weight assignment. A flexible feature collection is implemented which allows for collecting target specific labels from multiple training datasets. First tests were carried out using four different classes (*Handwriting, Printed, Image, Noise*). However, the system is capable to classify any other (visual) elements if they are visually dissimilar and properly trained.[4]

Similarly to the local orientation estimation, a multi-label graph-cut will be used to improve the labeling results with respect to local neighborhoods.

Figure 3 shows the component labeling. The lines represent the super-pixel's local orientation. Note that the local orientation substantially changes in the image area if no graph-cut is applied. These noisy estimations are harmonized after applying the graph-cut (right). The right image shows first test results if four different classes are trained. Super-pixels that are falsely labeled as *handwritten*, result from a strong handwriting prior within the training data.



mond, Wash., and probed visionaries coast to coast. In all, I talked to more than 100 people and heard dozens more at industry gatherings. My journey went from the ridiculous to the sublime: One morning, I watched a startup pitch a high-tech contraption to make fertilizer from cow chips. That afternoon, I visited Xerox' legendary Palo Alto Research Center (PARC) and saw computer chips whose microscopic circuits were literally printed using inkjet printers—a breakthrough that could produce disposable cell phones or ultracheap PCs.

Did I find the Next Big Thing? Truth be told, not exactly. But maybe, just maybe, I unearthed

Figure 3: Detail of a sample image published with the Page Segmentation Contest 2009 (left), local orientations *without* the multi-label graph-cut (middle), component labeling with four potential classes (right). The classes are: *Handwriting* (blue), *Printed* (yellow), *Image* (green), *noise* (gray).

5 Text-Line Segmentation

A preliminary text-line estimation is implemented which works on machine printed documents only. The super-pixels are therefore connected using Delauney triangulation. Since, we know the local orientation, the distances are weighted with each super-pixel's orientation. Hence, rather than using the Euclidean distance, the scalar product between the orientation vector and the edge vector is used. In the ideal case where two super-pixel centers are perfectly aligned and the orientation estimation is correct, the

[4] Here is a manual for training new classes: `https://github.com/TUWien/ReadModules/blob/master/manuals/Training%20Super%20Pixels.md`

distance becomes 0. While edges which are perpendicular to the local orientation have the Euclidean distance (the orientation vector is normalized). This methodology works correctly if text-lines need to be found within printed text. However, handwritten text is too complex to be tackled with a simple text-line detection like this (see Figure 5).

In order to improve the text-line segmentation, a tab-stop analysis is performed similar to that proposed by Ray Smith [5]. Figure 4 shows the results of the text-line segmentation on a printed document sample having skewed and multiple orientated text-lines. Colored polygons represent the text-lines' convex hulls while gray lines indicated the connected components.

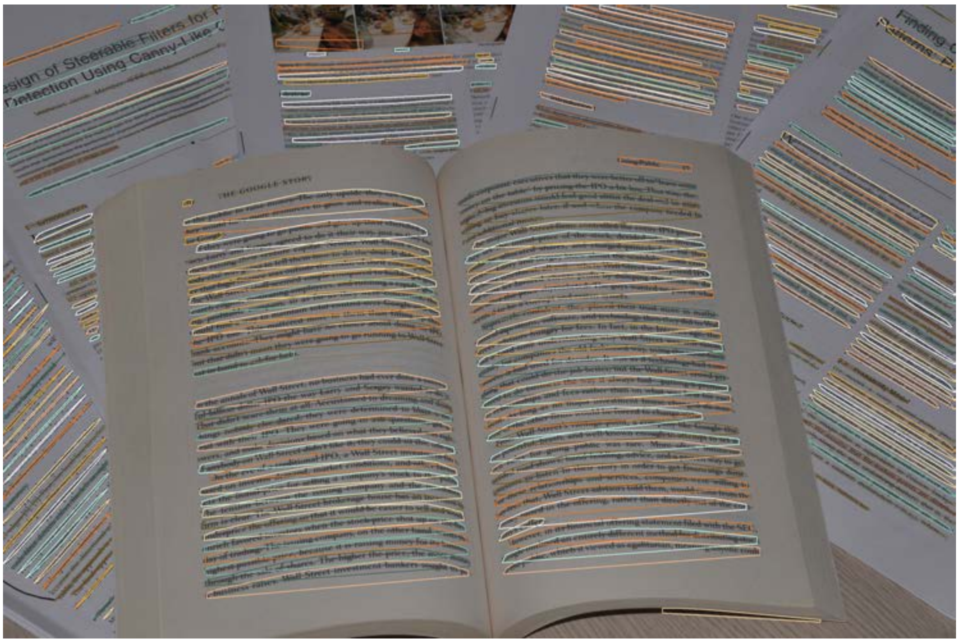


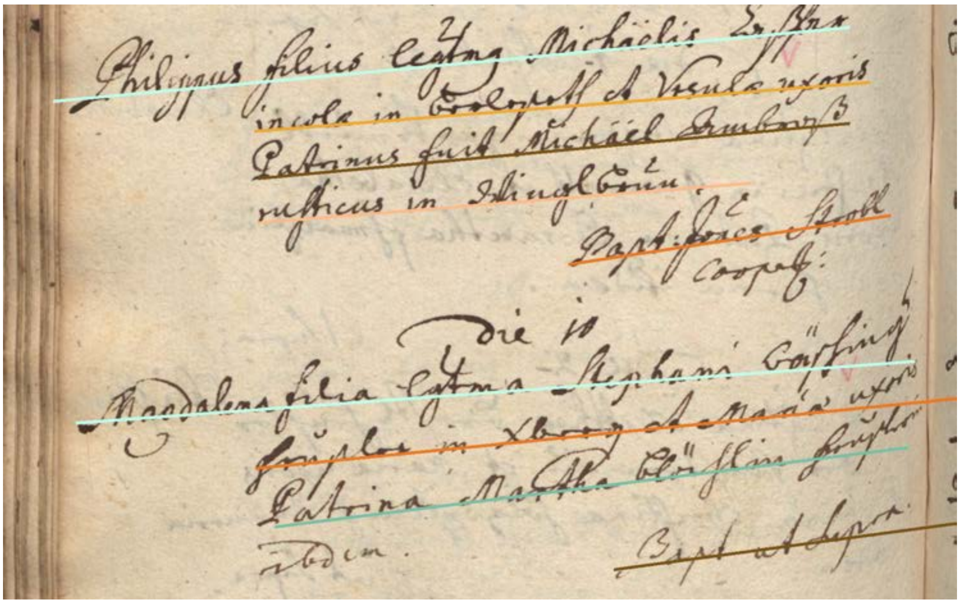
Figure 4: Text-line segmentation of a book having curved pages.

6 Evaluation

The basis for the method presented is a closed source in-house solution [6] which was evaluated on the past Handwriting Segmentation Contests. Table 1 gives an overview of its performance (CVL). In addition, its F-Score (FM) is compared with all participating methods on the last three Handwriting Segmentation Contests in Figure 6. The page segmentation (classification of elements) is evaluated on the ICDAR 2009 Page Segmentation Contest (see Figure 7).

7 Future Work

We are developing a new basic layout analysis module rather than using the existing in-house solution discussed in [6]. This is partly because of copyright and design issues.



Philippus filius legitimus Michaëlis [illegible]
incola in [illegible] et Ursula uxoris
Patrinus fuit Michaël [illegible]
rusticus in [illegible]
Bapt: Joannes Stroll
Cooperat:
Die 10
Magdalena filia legitima Stephani [illegible]
[illegible] in [illegible] et Maria uxoris
Patrina Martha [illegible]
ibidem.
Bapt. ut supra

Figure 5: Baseline visualization on a handwritten document from the baseline competition (*T_Freyung_003_0204*).

	M	o2o	DR	RA	FM		M	o2o	DR	RA	FM
CUBS	4,036	4,016	99.55	99.50	99.53	CUBS	1,626	1,589	97.54	97.72	97.63
ILSP-LWSeg-09	4,043	4,000	99.16	98.94	99.05	NifiSoft	1,634	1,589	97.54	97.25	97.40
CVL	4,034	3,977	98.59	98.59	98.59	**CVL**	1,633	1,583	97.18	96.94	97.06
PAIS	4,031	3,973	98.49	98.56	98.52	IRISA	1,636	1,578	96.87	96.45	96.66
CMM	4,044	3,975	98.54	98.29	98.42	ILSP-a	1,656	1,567	96.19	94.63	95.40
CASIA-MSTSeg	4,049	3,867	95.86	95.51	95.68	ILSP-b	1,655	1,559	95.70	94.20	94.95
PortoUniv	4,028	3,811	94.47	94.61	94.54	TEI	1,637	1,549	95.09	94.62	94.86

Table 1: Results of the ICDAR 2009 (left) [7] and ICFHR 2010 (right) [8] Handwriting Segmentation Contest.

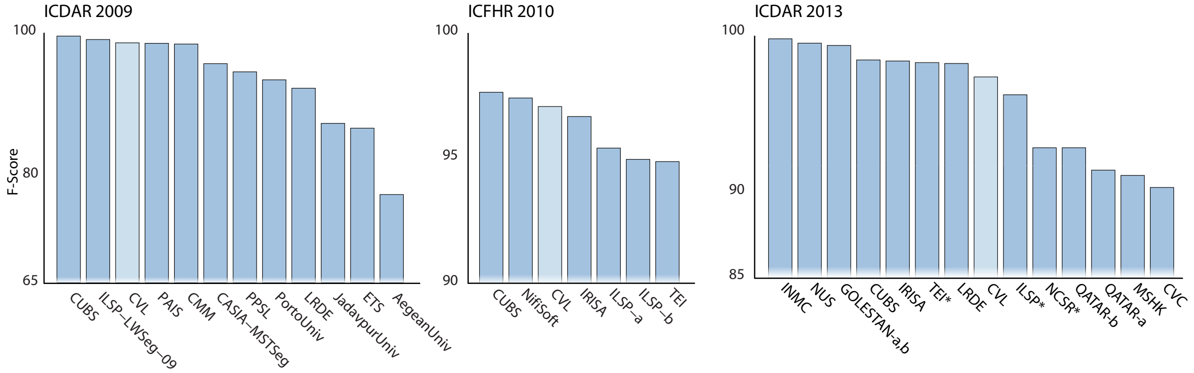


Figure 6: Results of the ICDAR 2009 [7], ICFHR 2010 [8], and ICDAR 2013 [9] Handwriting Segmentation Contest.

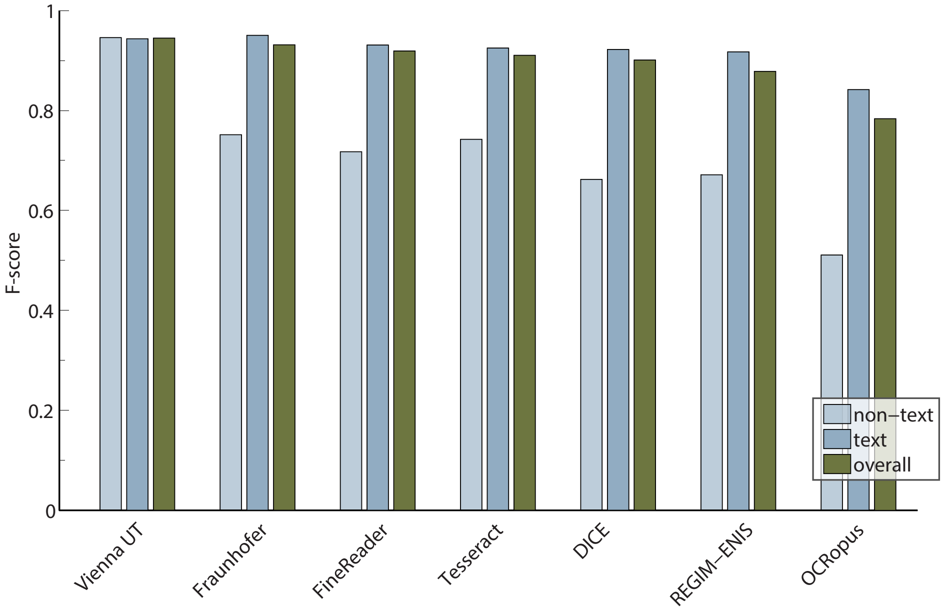


Figure 7: Comparison of CVL (Vienna UT) page segmentation with all participating methods of the ICDAR 2009 Page Segmentation Competition.

The existing solution is designed for modern documents while the proposed method targets additionally old archival documents. The key differences are:

	Old	New
CC Segmentation	Binarization	MSER
Skew	global	local (per CC)
Labeling	3 classes	flexible

That is why handwritten text-line segmentation and page segmentation (zoning) are not fully functional yet. It is planned to group the CCs using their class labels, locations, and lines (graphical lines or virtual tabstop lines). By these means a zoning will be established which allows other algorithms to perform their tasks on specifically labeled zones of a document image. Moreover, the text-line clustering has to be improved for handwritten archival documents.

References

[1] B. Su, S. Lu, and C. L. Tan, "Binarization of Historical Document Images Using the Local Maximum and Minimum," in *DAS '10: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems.* New York, NY, USA: ACM, 2010, pp. 159–166.

[2] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," in *Proceedings of the*

British Machine Vision Conference 2002, BMVC 2002, Cardiff, UK, 2-5 September 2002, 2002, pp. 1–10. [Online]. Available: http://dx.doi.org/10.5244/C.16.36

[3] Hyung Il Koo, “Textline Detection in Camera-captured Document Images using the State Estimation of Connected Components,” in *Transaction on Image Processing*, 2016.

[4] Tin Kam Ho, “The Random Subspace Method for Constructing Decision Forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998. [Online]. Available: http://dx.doi.org/10.1109/34.709601

[5] Raymond W. Smith, “Hybrid Page Layout Analysis via Tab-Stop Detection,” in *10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26-29 July 2009*, 2009, pp. 241–245. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2009.257

[6] Markus Diem, Florian Kleber, and Robert Sablatnig, “Text Line Detection for Heterogeneous Documents,” 2013, pp. 743–747.

[7] Basilios Gatos, Nikolaos Stamatopoulos, and Georgios Louloudis, “ICDAR 2009 Handwriting Segmentation Contest,” in *ICDAR*, 2009, pp. 1393–1397.

[8] ——, “ICFHR 2010 Handwriting Segmentation Contest,” in *ICFHR*, 2010, pp. 737–742.

[9] Nikolaos Stamatopoulos, Basilis Gatos, Georgios Louloudis, Umapada Pal, and Alireza Alaei, “ICDAR 2013 Handwriting Segmentation Contest,” 2013, pp. 1402–1406.